Thomas Bennett
CS 4395.001 - UTD Spring ‘23
Portfolio Assignment 0

Overview of NLP

Natural Language Processing (NLP) is how computers understand human speech. Classical NLP used grammar rules to model human speech. Modern NLP uses statistical models created with machine learning algorithms to predict what the next word should be in a body of text. NLP technologies are concerned with the creation of human-like speech (natural language generation) as well as understanding human speech (natural language understanding).

Thanks to technologies such as ChatGPT, DALL-E, Alexa, and Cortana, NLP is a hot topic in the news lately. ChatGPT can produce bodies of text strikingly similar to what a high school student might write when given a writing prompt. DALL-E uses NLP to produce striking visuals based on human language descriptions of an image. Alexa, Cortana, and other virtual assistants use NLP to understand what you are telling them to do.

The earliest natural language processing systems consisted of grammars. A grammar is a set of rules for generating sentences of a language. This early period of NLP development is sometimes referred to as symbolic NLP.

Examples of symbolic NLP technologies:

- STUDENT: capable of solving algebra word problems
- ELIZA: a simulated therapist, and early example of a chatbot
- LIFER/LADDER: a natural language query database system for the U.S. Navy
- Racter: randomly generated correctly formed English sentences, used to produce a book, *The Policeman’s Beard is Half Constructed*

The era of statistical NLP began in the 1980s. As computing power increased and powerful machines became more widely available in research institutions, NLP researchers began using probabilistic models developed using machine learning algorithms to generate and understand text. IBM’s Watson showcased the power of statistical NLP when it famously defeated two of the best human *Jeopardy!* contestants. Watson’s systems used statistical NLP to understand the provided clue, and to provide the answer stated as a question.

Deep learning NLP uses neural networks rather than large feature vectors to produce language models. However, the very large neural networks needed to produce good language models are

prohibitively expensive. Similar to statistical NLP, deep learning NLP requires a large amount of training data to produce a good language model. IBM’s Watson and other contemporary NLP technologies now use neural networks as part of their deep learning systems.

NLP has greatly benefited from advances in artificial intelligence. Machine learning, a subfield of artificial intelligence, provides learning algorithms which can improve themselves as they receive input data. NLP leverages this ability to learn to create language recognition and generation models.

My personal interest in NLP is growing as I learn more about machine learning and as NLP technologies mature. I have little doubt that NLP will be a large part of future technologies, such as operating systems. Virtual assistants already exist in most operating systems, so it is logical to assume that they will continue to be present and will receive upgrades. I have some concerns relating to NLP though. NLP technologies use large bodies of text as training data. The largest body of human-produced text is the Internet. The Internet is not without biases. Since the development of these technologies is largely an opaque process, I do not know whether these biases are being examined and perhaps avoided. A question I will be interested in answering throughout CS 4395 is: “how can otherwise innocent NLP technologies be used to harm people?” Technologies are amoral, therefore they may be as capable as helping as harming. To protect people from harm, we must first understand what harm is possible.

Thomas Bennett
CS 4395.001 - UTD Spring ‘23
Portfolio Assignment 0

Bibliography

Artsrouni, G. (n.d.). *History of natural language processing*. Wikipedia. Retrieved January 28, 2023, from https://en.wikipedia.org/wiki/History_of_natural_language_processing

Hendrix, G. G., Sacerdoti, E. D., Sagalowicz, D., & Slocum, J. (1978, June). Developing a natural language interface to complex data. *ACM Transactions on Database Systems*, *3*(2), 105-147. https://doi.org/10.1145/320251.320253

Henrickson, L., Roggenbuck, S., Spinosa, D., & Murray, J. (2021, April 4). *Constructing the Other Half of The Policeman's Beard*. Electronic Book Review. Retrieved January 28, 2023, from http://electronicbookreview.com/essay/constructing-the-other-half-of-the-policemans-beard/

IBM Watson. (n.d.). Wikipedia. Retrieved January 28, 2023, from https://en.wikipedia.org/wiki/IBM_Watson

IBM Watson Studio - Deep Learning. (n.d.). IBM. Retrieved January 28, 2023, from https://www.ibm.com/cloud/watson-studio/deep-learning

LIFER/LADDER. (n.d.). Wikipedia. Retrieved January 28, 2023, from https://en.wikipedia.org/wiki/LIFER/LADDER

Racter. (n.d.). Wikipedia. Retrieved January 28, 2023, from https://en.wikipedia.org/wiki/Racter

Searle, J. (n.d.). *Natural language processing*. Wikipedia. Retrieved January 28, 2023, from https://en.wikipedia.org/wiki/Natural_language_processing

What is Natural Language Processing? (n.d.). IBM. Retrieved January 28, 2023, from https://www.ibm.com/topics/natural-language-processing

What is Natural Language Processing? (n.d.). Google Cloud. Retrieved January 28, 2023, from https://cloud.google.com/learn/what-is-natural-language-processing